



10GE Testing in ESnet

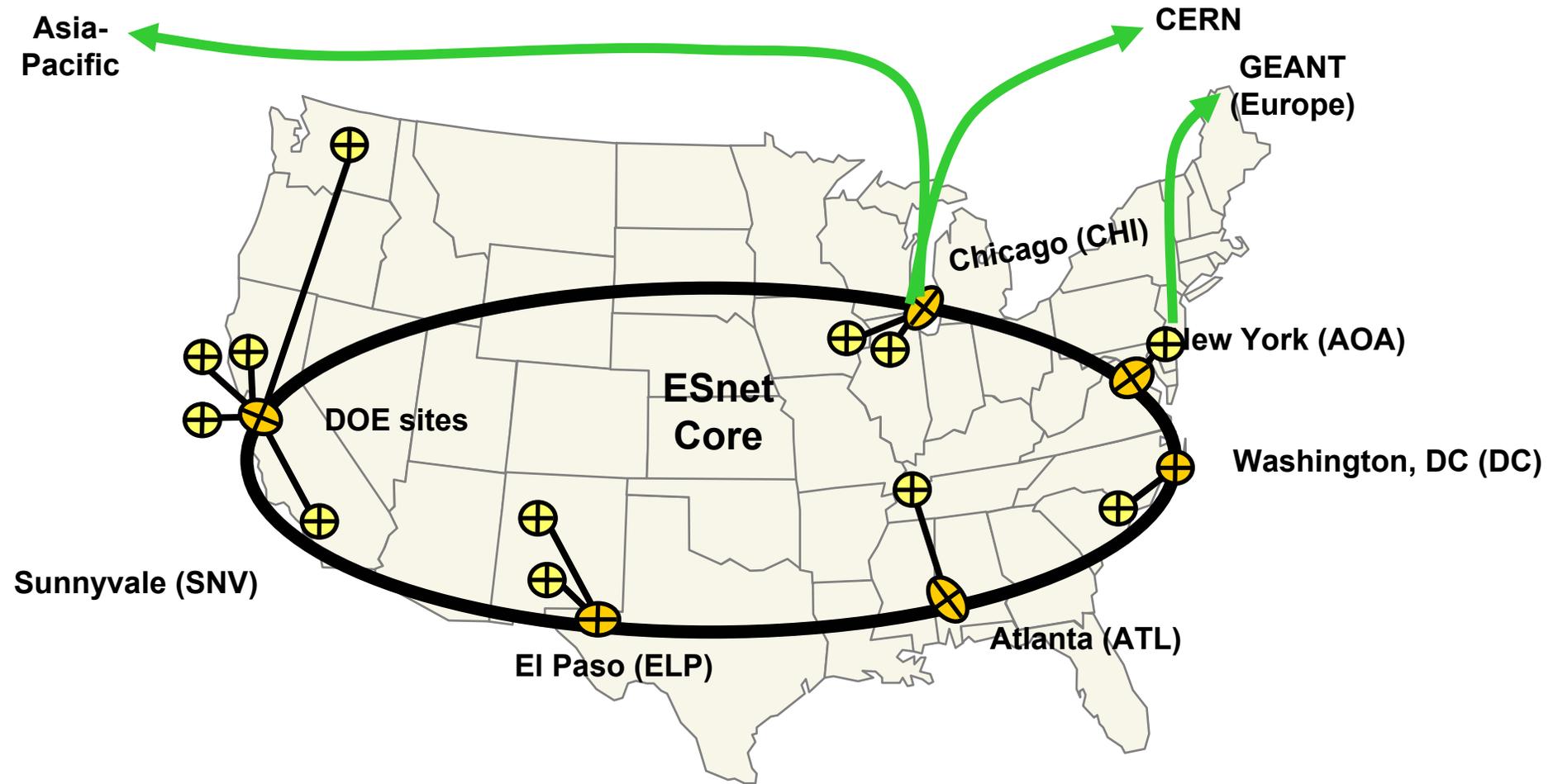
Joe Metzger

metzger@es.net

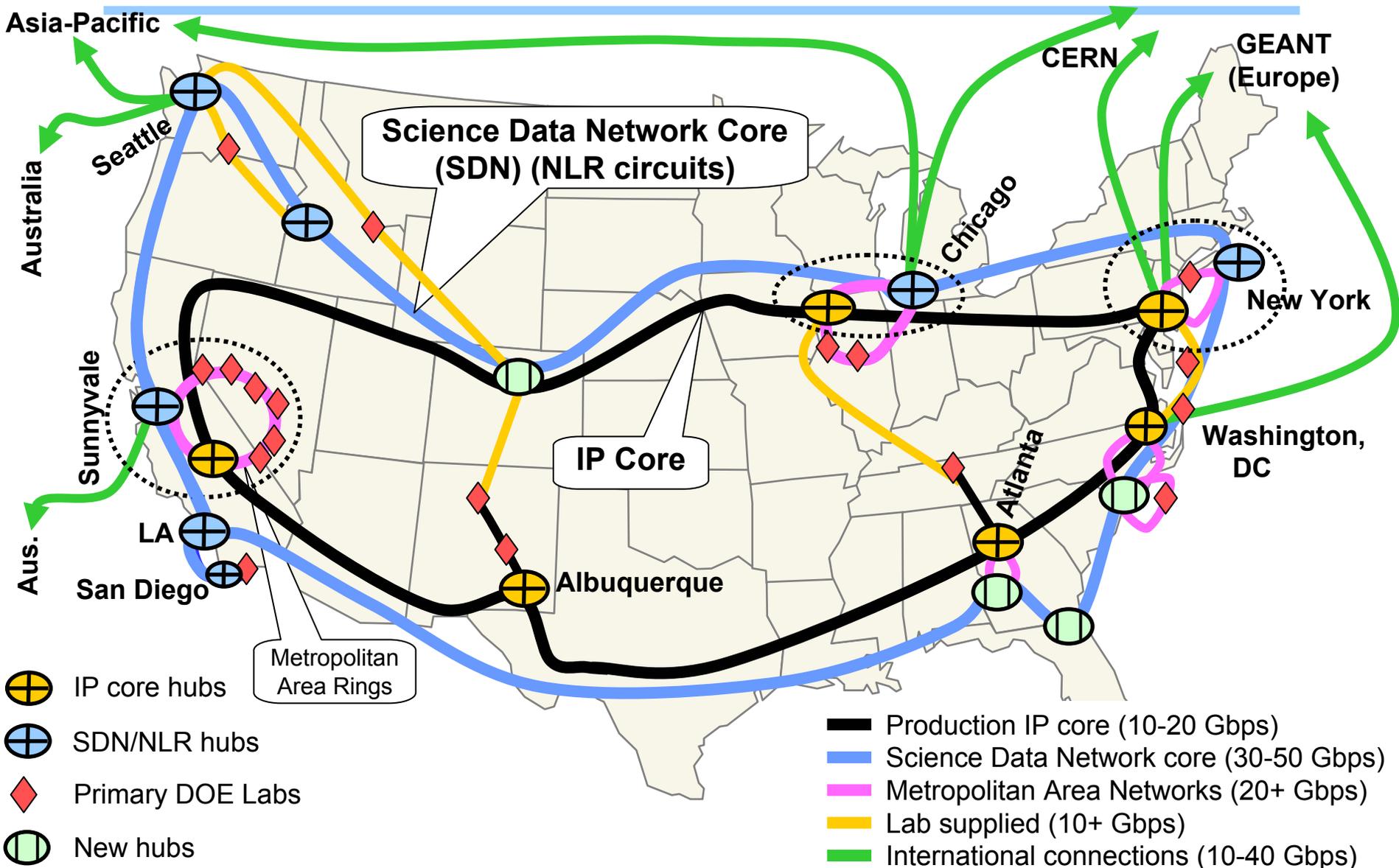
SC 05



Evolution of ESnet: Past



Evolution of ESnet: Future



New 10 GE Circuits in ESnet

We are 2/3^{rds} of the way through a plan to install more than 30 new 10GE circuits within a 12 month period.

- May – October 05
 - 16 circuits composing the San Francisco Bay Area MAN
 - 3 circuits as part of the SDN network: San Diego to Sunnyvale, Sunnyvale to Seattle, and a connection between POPs in Sunnyvale.
- November 05 – Spring 06
 - 3 circuits in New York to Brookhaven
 - 9 circuits in Chicago to FERMI and Argonne

SONET is Nice

- Performs continuous quality checks.
- Tracks both error counts and errored seconds.
- Tracks errors at multiple layers.
 - Section Checks (between repeaters)
 - Line Checks (between muxes)
 - Path Checks (between routers)
- Maintains lots of useful counters
 - LOS/LOF – Loss of Signal
 - Bit Error Rates
 - PLL for clocking problems
 - RDI – Remote Defect Indications
 - Remote Error Indications
- Crossing error thresholds generates alarms.

Ethernet is more Challenging

- No error detection when idle.
 - Defects can't be identified until data is sent.
- All signals fall into one of the following categories:
 - Valid Packets
 - Runts
 - Giants
 - Bad CRC
 - Jabbers

ESnet Acceptance Tests

- When is a circuit ‘good enough’?
- How many packets can you lose in 24 hours if a 10GE is running at line speed?
 - 0
 - 100
 - 1000
 - 2000

10GE Standard Bit Error Rates

- The 10GE standard specifies a BER of $10e-12$.¹
- Convert to a Frame Loss Rate²

FLR = Frame Loss Rate

BER = Bit Error Rate

N = Bits in Frame (9K IP +26 Bytes Ethernet Header)

$$\text{FLR} = 1 - (1 - \text{BER})^N$$

$$\text{FLR} = 1 - (1 - 10e-12)^{72,280}$$

$$\text{FLR} = 7.2e-7$$

1. 802.3ae clause 52.
2. Conversion formula from page 89 "Gigabit Ethernet" by Rich Seifert, Addison Wesley 1998.

ESnet BAMAN Acceptance Tests

- **Saturation Test**
 - The circuit shall be saturated with a demonstrated bandwidth over 95% of the link capacity for 5 minutes.
 - In situations where the integrity of the counters used to compute utilization are suspect, confirmation should be made by using a second set of counters.
 - This test is to assure that circuit that has been delivered doesn't have any internal bottlenecks that would prevent it from running at capacity.
- **Loss Test**
 - Greater than 50 Terabytes shall be transferred across the link in each direction with an error rate of less than one in **10e8 packets**. (1000 Packets)
 - This test is to assure that the line is running clean.

10 GE Performance Test Systems

- Tyan S2895A2NRF
Dual AMD 252 Opteron CPUs (2.6 Ghz)
2GB DDR400 ECC/Reg. Memory
120GB 7200RPM 8MB Buffer SATA Drive
- Neterion (S2IO) 10GE X-Frame PCI-X NIC
- Linux Fedora Core 3 2.6.10

Performance Tester Tuning

- 9K MTU
- Disable iptables
- `setpci -d '17d5:*' 62=15 LATENCY_TIMER=FF`
- SYSCTL
 - `net.ipv4.tcp_timestamps = 0`
 - `net.ipv4.tcp_sack = 0`
 - `net.ipv4.tcp_rmem = 20000000 20000000 10000000`
 - `net.ipv4.tcp_wmem = 20000000 20000000 10000000`
 - `net.ipv4.tcp_mem = 20000000 20000000 10000000`
 - `net.core.rmem_default = 1048754`
 - `net.core.wmem_default = 1048754`
 - `net.core.optmem_max = 1048754`
 - `net.core.netdev_max_backlog = 300000`
 - `net.core.wmem_max=20000000`
 - `net.core.rmem_max=20000000`

Cisco 6509 Counter Review

- CRCs Standard sized packets with bad CRCs
- Giants Large Packets
- Jabbers Large Packets with bad CRCs

- 6509 ACL counters don't count packets forwarded or dropped in hardware, only process switched packets.

- Interface counters also count miscellaneous packets.
 - CDP, SPDU, OSPF, PIM, ARP, SNMP etc.

Linux Counters

- Ifconfig RX and TX packets & Bytes
 - Average packet size = 34K ???
- “ethtool -S eth2” reports Good values for:
 - Tmac_tcp
 - Rmac_tcp
 - Tmac_data_octets
 - Rmac_data_octets

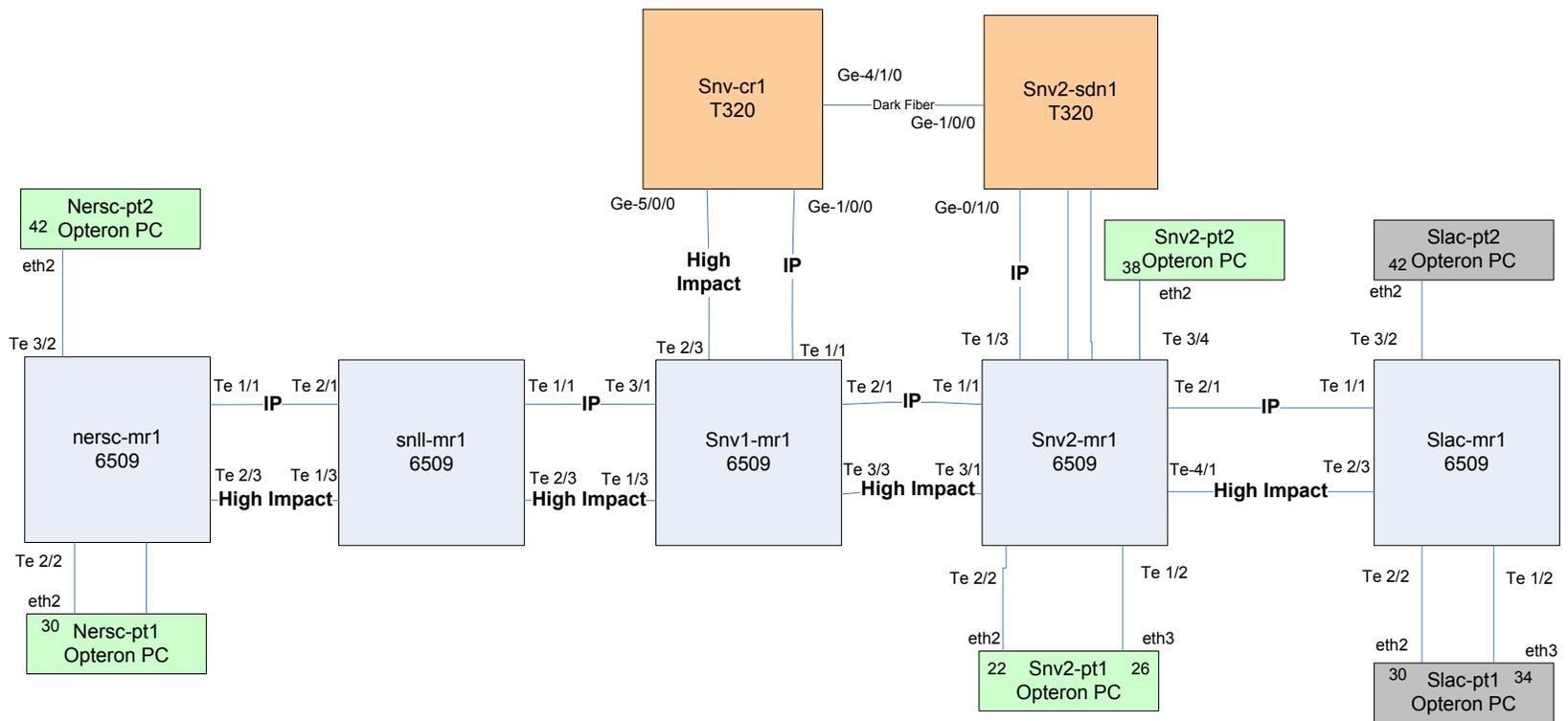
Juniper Firewall Filters

- They work!
- They can be used to check other counters
- They can help you figure out where packets were lost.

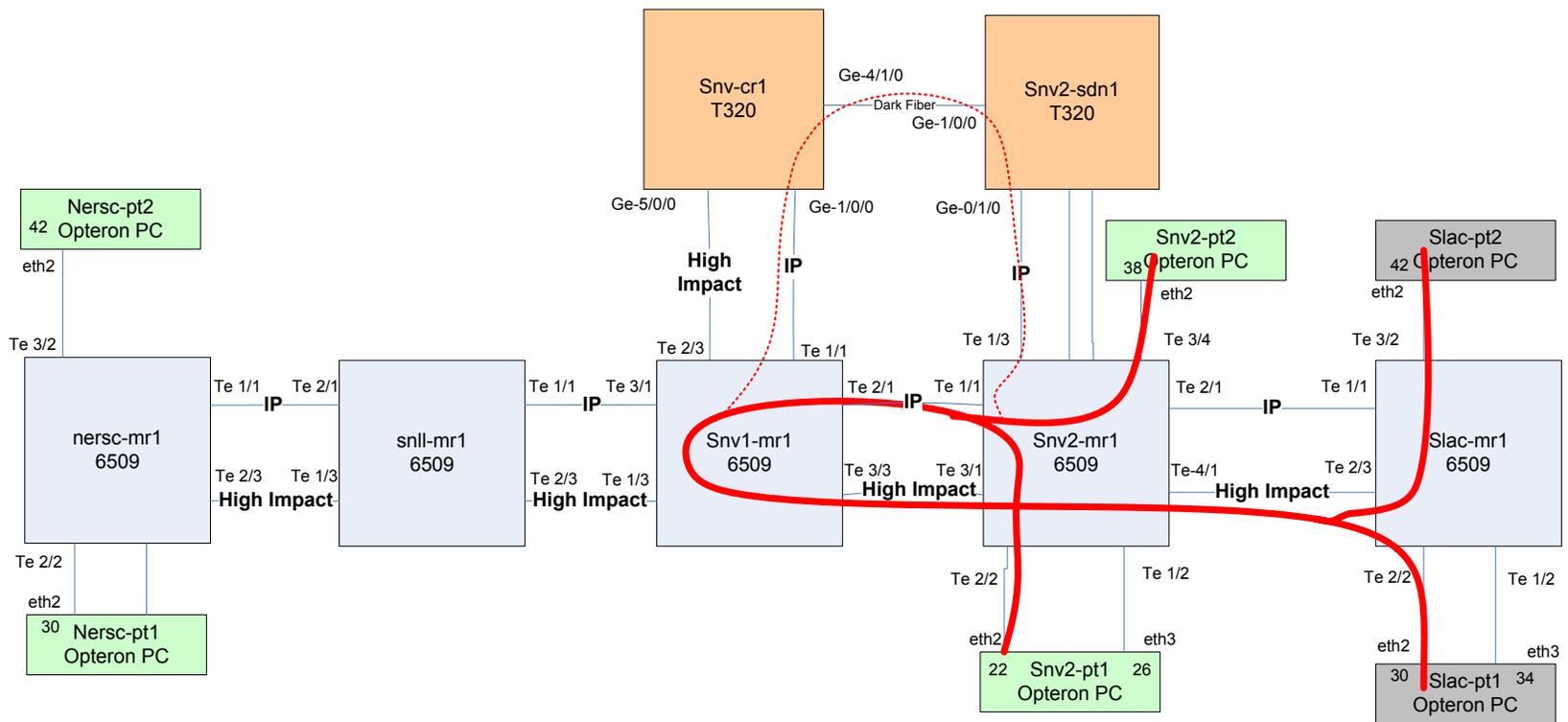
Loss Testing Procedure

- Configure routing or layer 2 paths to force test streams over the test links.
- Record all counters on routers and testers.
- Run IPERF tests recording output.
- Record all counters again
- Check results relying on Linux ethtool TCP counters or firewall filters.
- Double check for sanity using interface counters to confirm that traffic flowed over test link.

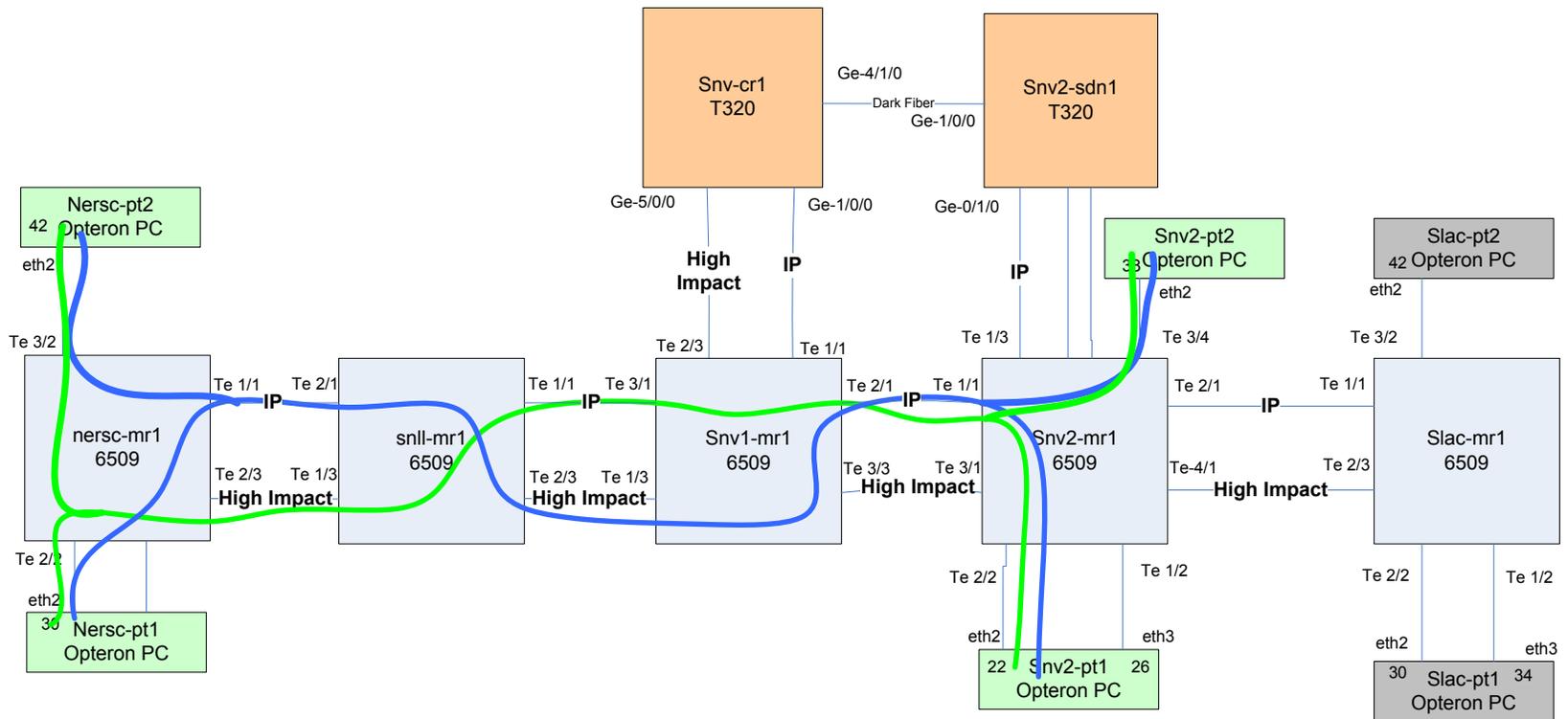
Early Bay Area MAN Configuration



Circuits Between SNV1 (Qwest) & SNV2 (Level3)



Circuits between SNV2 & NERSC via LLNL/SNLL)



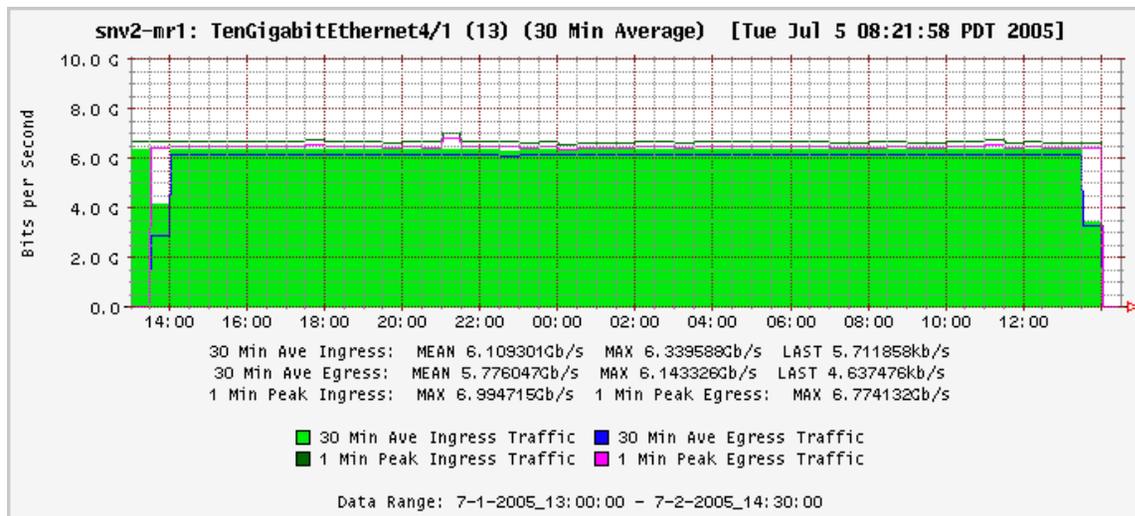
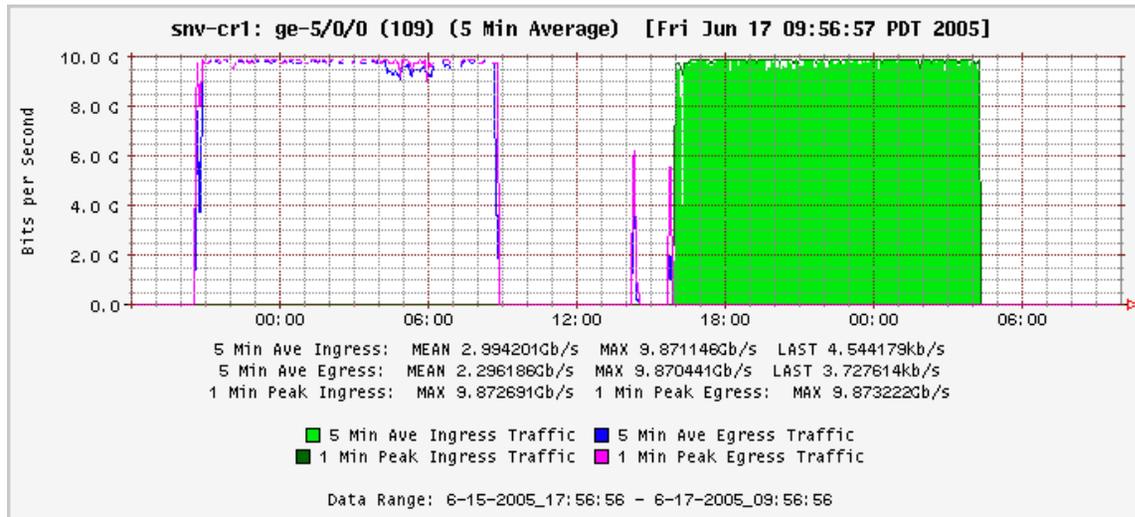
Test Results

Link	Saturation Test	Loss Test	
	Rate Achieved (> 9.5 Gbps)	Bytes Transferred	Packets Lost
1. SNV2 SLAC	9.838 Gbps	107.5 TB	0
2. SNV2 SLAC SDN	9.962 Gbps	134.7 TB	1
3. SNV-SNV2	9.962 Gbps	107.5 TB	0
4. SNV-SNV2 SDN	9.962 Gbps	107.5 TB	0
5. SNV-LLNL	9.757 Gbps	136.5 TB	0
6. SNV-LLNL SDN	9.752 Gbps	136.5 TB	0
7. LLNL-NERSC	9.752 Gbps	136.5 TB	0
8. LLNL-NERSC SDN	9.757 Gbps	136.5 TB	0

Test Results 2

Link	Saturation Test	Loss Test	
	Rate Achieved (> 9.5 Gbps)	Bytes Transferred	Packets Lost
9. SLAC-LBL	9.781 Gbps	120.7 TB	2
10. SLAC-LBL SDN	9.773 Gbps	114.4 TB	1
11. NERSC-JGI	9.947 Gbps	124.8 TB	180*
12. NERSC-JGI SDN	9.943 Gbps	131.3 TB	84+
13. LBL-JGI	9.947 Gbps	124.8 TB	180*
14. LBL-JGI SDN	9.943 Gbps	131.3 TB	84+
SNV2 - SDSC	9.878 Gbps	113.9 TB	24
SNV2 – SNV1	9.962 Gbps	107 TB	0

Sample Result Graphs



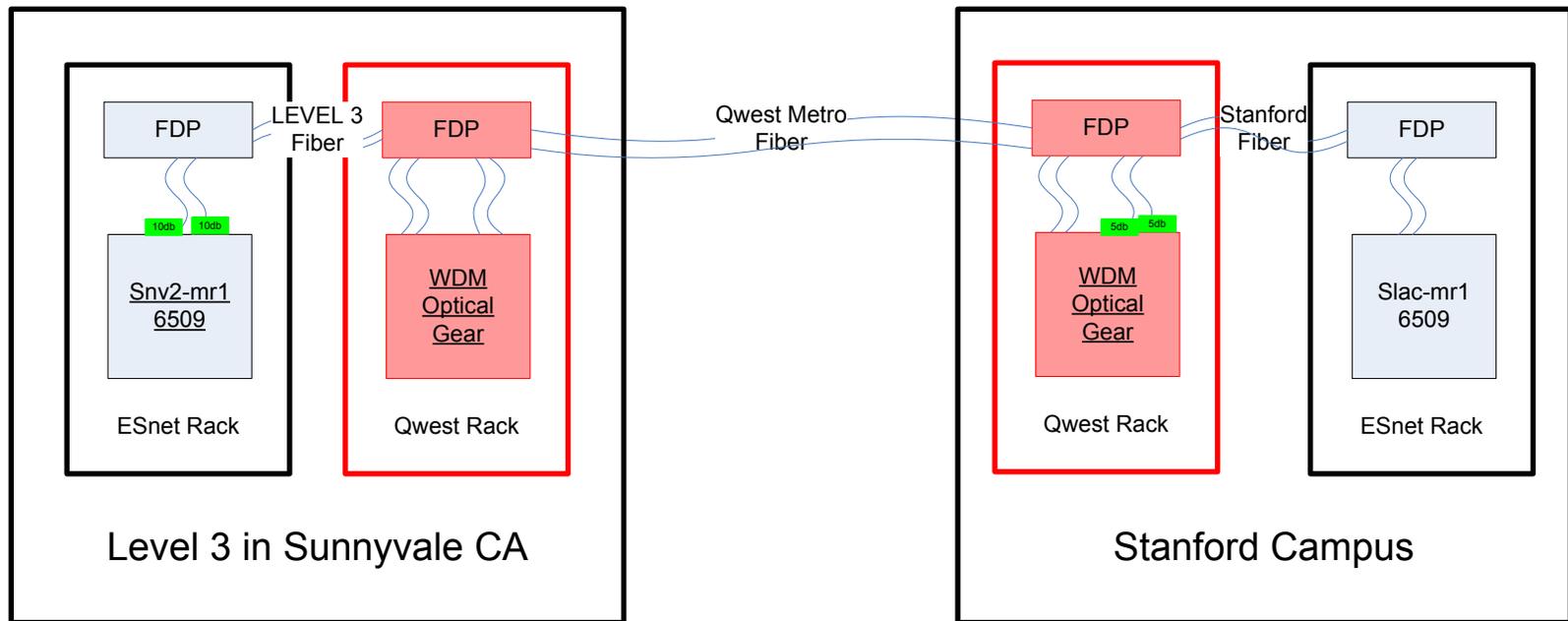
Challenges

- Lots of Variables
- Dirty Connections
- Power Levels
- Cisco IOS Versions
- TCP Tuning
- 1 Armed Routing Bugs
- How many errors will we miss?

SNV2 to SLAC Circuit

A Simple example:

- About 40 fiber connections, which one is dirty?
- Is the receive power in the middle of the specification on every interface?



SNV2 – SLAC

High Impact Circuit

- 1 Input Queue drop per Second at 9.9Gbps
 - Cisco Bug CSCeg62365
 - Replaced 12.2(18)SXD3 with 12.2(17d)SXB8
- 2.7 Packets lost per Minute at 9.9Gbps.
 - Added a 5 db pad
- ~9 packets lost per Hour at 9.9 Gbps
 - (2.75 packets lost per hour at 6.1 Gbps)
 - Cleaned and reseated fibers at SNV2
- ~.75 Packets lost per hour at 6.1 Gbps
 - Cleaned and reseated fibers at SLAC
- 1 Packet Lost in 24 Hours at 6+ Gbps.

TCP Tuning

- Poor performance when window size is too small.
 - Can't keep pipe full.
- Poor Performance when window size too big.
 - Drops 1 packet
 - Transmits at 10's of Kbps
 - RTT seen by TCP goes up to 30 seconds.
 - Recovers
 - Repeat
- Newer Kernels may fix some of the problems.

TCP Tuning

- What size window should you use with 2 competing streams over a 10GE link?
 - SLAC to SNV2 0.63 ms RTT
 - 2 MB windows -> 9.96 Gbps
 - SLAC to SNV2 via SNV 1.21 ms RTT
 - 2 MB windows -> 9.96 Gbps
 - NERSC to SNV2 1.97 ms RTT
 - 2.00 MB window -> 9.33 Gbps
 - 2.05 MB window -> 9.58 Gbps
 - 2.1 MB window -> 8.60 Gbps
- I was unable to make 2 competing TCP streams saturate a link from San Francisco to San Diego (13 ms).
 - Patching or replacing the TCP stack should resolve this.
 - Resorted to UDP and looping packets through the link more than once.

1 Armed Routing Bug

- Testing Sunnyvale to San Diego SDN link required sending packets between testers in SF Bay area down to San Diego and back over 1 link.
- The Cisco 6509 was unable to forward more than 1 Gbps of traffic when 1 armed routing on a single VLAN using a static route (or via an MPLS LSP).
- The box can forward packets in hardware between 2 vlans using a static route.

Undetected Errors?

- We are only looking at lost packets and CRC errors.
- Research has shown rates of errors undetected by link CRC's and TCP checksums ranging from one in 16 million to 10 billion packets.
 - When the CRC and TCP checksum disagree
 - Stone & Partridge, SIGCOMM 2000, pgs 309-313
 - <http://portal.acm.org/citation.cfm?doid=347059.347561>
 - 16 Million 9K packets can be sent in less than 2 minutes on a 10GE link.
 - 10 Billion 9K packets can be sent in less than 24 hours on a 10GE link.

Summary

- 10 Gigabit Ethernet
 - A challenge in the metro and wide area.
 - Some circuits perform flawlessly from the start.
 - Others require extensive testing & diagnosis.
 - Comprehensive stress testing is essential!
- ESnet is 2/3rds of the way through a plan to install more than 30 new 10GE circuits within a 12 month period.
 - The SF Bay Area MAN is up and running.
 - Circuits in Chicago and Long Island are in progress.
 - We have started to deploy a second backbone to increase throughput and reliability